行政院國家科學委員會專題研究計畫成果報告

以 XML 為基礎之分散式模糊知識管理系統研究 (I)

Study on XML-based Distributed Fuzzy Knowledge Management System (I)

計畫編號：NSC 91-2413-H-032-007
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：林信成　　執行機構：淡江大學資圖系

壹、中文摘要

本研究結合模糊理論與全文標示兩種不同的作法，達到提升檢索系統回現率與精確率之目的。我們以電子新聞與傳統圖書作為實驗對象，主要研究成果有二：(1) 針對圖書資料內容進行軟性的模糊分類，依據模糊理論的概念賦予每筆資料在各類別上的模糊歸屬函數值，以便檢索系統能依據圖書資料之歸屬度進行模糊關聯搜尋，使得檢索方式更具彈性，亦可提供較佳的回現率；(2) 針對新聞資料內容以 XML 格式進行全文標示，利用 XML 進行語意定義與描述，使資料內容具備自我描述性，以便檢索模組進行較精確的語意檢索，使檢索結果能夠達到較高的精確率，以更符合使用者的資訊需求。

關鍵字：模糊理論、模糊分類、全文標示、XML、Fuzzy、知識管理

貳、緣由與目的

一、模糊理論

模糊理論實際上是模糊集合（Fuzzy Set）、模糊關係（Fuzzy Relation）、模糊邏輯（Fuzzy Logic）、模糊控制（Fuzzy Control）、模糊量測（Fuzzy Measure）… 等理論之泛稱[1]，是一門用以將模糊概念量化的學問，起源於 1965 年扎德（L.A. Zadeh）教授所發表的著名論文－「模糊集合」[2]。模糊理論以模糊集合為基礎，以研究不確定事物為目標，接受模糊現象存在的事實，根據不清晰訊息，透過近似推理（Approximation Reasoning）[3]過程而得到正確結果，這與人腦「過程模糊，結論清晰」的思維方式極其類似，因此已被廣泛的應用於各種不同領域的智慧型系統中[4]。

傳統明確集合（Crisp Set）的特徵函數（Characteristic Function）採用非 0 即 1 的二分法，而模糊集合的基本精神則是將其擴展成由 0 至 1 的任何值，稱為歸屬函數（Membership Function），當一個元素屬於某集合的程度越大時，其歸屬程度就越接近於 1，否則越接近於 0。若以 X 代表論域（Universe of Discourse），A 代表 X 中的任一離散型模糊集合，則可將 A 表示成如下型式：

$$A=\sum_{i=1}^{n}\mu_i / x_i = \mu_1 / x_1 + \mu_2 / x_2 + \ldots + \mu_n / x_n$$

其中，$x_i \in X$ 為 X 中任一點。藉由歸屬函數對模糊概念進行量化之後，便可以利用精確的數學方法進行模糊資訊的分析和處理了。

二、全文標示

Web 是網路時代最重要的資訊傳播平台，而 HTML 則是發行 Web 電子文件的標準規範。然而，HTML 擅長於版面編排與外觀格式，對於文件結構的規範及內容語意的描述則相對不足；XML 的誕生正好提供了一個可行的解決方案，彌補 HTML 之短。XML 自 1998 年日 2 月 10 正式標準[5]發佈至今，歷經五年多的發展[6]，已經成為一個陣容龐大的技術家族：DTD 和 XML Schema[7]用以定義文件的結構；CSS[8]和 XSL/XSLT[9]分別作為呈現文件版面和轉換文件格式之用；DOM[10]是剖析文件時的標準物件模型；RDF[11]則作為 Metadata 之整合框架；Nmaespace[12]是一致性名稱識別機制；以及各種衍生的應用語言如 MathML[13]、SVG[14]、PNG[15]、SMIL[16] … 等，再加上由各個不同組織基於 XML 所發展出適用於各行各業的應用語言將近千種[17]，真可謂族繁不及備載。至於近期 XML 的研究則逐漸朝向智慧型的 Web 語意網（Semantic Web）[18]和 Web 知識體（Web Ontology）[19]的方向發展，使得 Web 系統與文件皆愈來愈智慧化。

以 XML 進行全文標示（Full-Text Markup）的電子文件不但內容和外觀分離，且具備了結構性、整合性和自我描述性等重要特色，不但能有效解決目前網路上電子文件的亂象，更有助於開創智慧型電子出版的新契機。

參、結果與討論

一、電子新聞管理系統

本研究所設計之電子新聞管理系統，可分為前端子系統與後端子系統兩部分，如**錯誤! 找不到參照來源。**所示。若依功能劃分則可區分為五大部分，分別是：(1) 呈現模組（Presentation Module）；(2) 檢索模組（Searching Module）；(3) 編輯模組（Editing Module）；(4) 管理模組（Management Module）；(5) 資料庫（Database）。

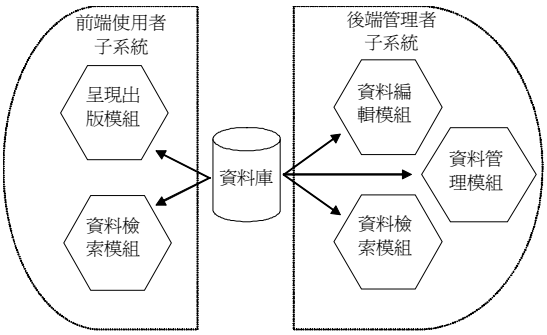


圖 1 系統功能模組

主要功能說明如下：

(1) 呈現模組

本模組負責將資料內容呈現給使用者，主要負責呈現使用者經過檢索後所需閱讀的資料。由於本系統內之文件資料儲存方式以 XML 為主，因此本模組具備解讀 XML 文件的功能。此模組透過 DSO 以及 DOM 來解讀 XML 文件，並結合該資訊所需之相關功能及超連結，加以包裝、排版，只要相容於本系統之 DTD 規範的 XML 文件，皆可透過此模組呈現內容。

(2) 檢索模組

本模組提供新聞資料檢索功能，主要由「模糊檢索引擎」與「語意檢索引擎」所構成：

(a) 模糊檢索引擎：提供新聞分類的模糊檢索功能，使用者可決定欲查詢之新聞資料其所屬分類的歸屬函數值(0 至 1)，模糊檢索引擎針對所有新聞資料錄之歸屬函數值，在其對應的模糊集合當中依據使用者的檢索條件進行搜尋，使得其檢索的方式更具彈性，以提供較佳的回現率。

(b) 語意檢索引擎：傳統的搜尋引擎僅提供資料欄位的檢索功能，並未提供針對內容語意上之特定目標加以檢索，如人名、地名等，本模組除提供傳統的欄位與關鍵字詞檢索功能之外，藉由 XML 具備資料自我描述之特性，可針對新聞內容的人、事、時、地、物等內文語意加以檢索，提高檢索結果的精確率。

(3) 編輯模組

本模組提供管理者編輯新聞文件內容。透過本模組可將新聞內容編輯成符合系統之 DTD 規範的 XML 文件，並針對新聞內容給予人、事、時、地、物不同的語意標示，以及可針對新聞系統內各新聞分類給予歸屬函數值，以利提供檢索模組進行內文語意檢索以及模糊分類檢索等功能。另外經由 XML 文件內容與呈現資料分離的特點，同一份文件可選擇不同樣式來排版。

(4) 管理模組

本模組提供管理者異動／修改資料。透過系統之資料管理模組，可針對新聞資料做新增、刪除、修改等各項異動，並且內建資料檢索功能，以利於管理者於後端進行資料維護時提供便捷的資料查詢功能。

(5) 資料庫

關聯式資料庫是目前最通用的資料庫系統，因此，在本研究中我們採用關聯式資料庫系統。我們依據上一單元所分析的模糊分類模式，可以很容易的將模糊關係矩陣轉換為關聯式資料表的模糊分類欄位，如表 1 所示，並給予歸屬函數值，使得檢索模組可藉以判別新聞所屬之模糊類別。

	c_1	c_2	…	c_m
d_1	μ_{11}	μ_{12}	…	μ_{1m}
d_2	μ_{21}	μ_{22}	…	μ_{2m}
⋮	⋮	⋮	…	
d_n	μ_{n1}	μ_{n2}	…	μ_{nm}

表 2 模糊關聯式資料表

此系統的運作分為使用者與管理者兩種流程。使用者可經由資料檢索模組，輸入欲查詢之條件(可選擇搜尋模式：模糊檢索或語意檢索)，如圖 2 所示，隨後資料檢索模組至資料庫中搜尋所需資料。檢索模組的檢索工作完成後，由資料庫所取出之資料後交由資料呈現模組，依照系統所訂定之 DTD 來確認 XML 資料的合法性，隨後將呈現給使用者。呈現的 XML 資料透過 DSO 以及 DOM 的解讀，可在不更改原始新聞內容之下，給予各種排版樣式，如圖 3 所示。

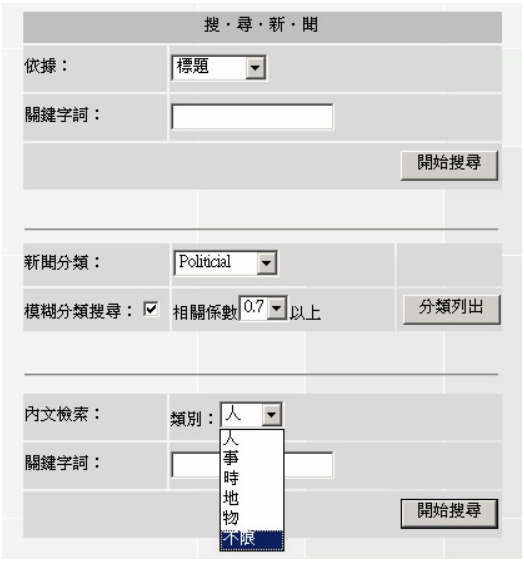
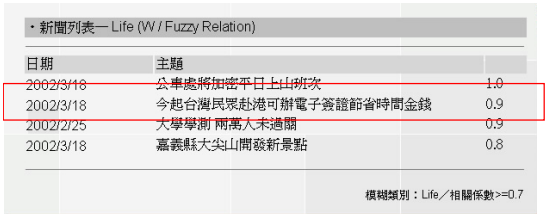


圖 2 資料檢索模組

圖 3 各種排版樣式

(1) 模糊檢索結果

模糊檢索引擎的目的在於提供較佳的回現率。以本系統資料庫內之新聞『今起台灣民眾赴港可辦電子簽證節省時間金錢』為例，由於內容描述香港政府之旅遊發展局等相關單位有鑑於台港旅遊人次繁多，簽證手續繁雜，因此推出新的簽證方式以節省時效。在實質上可歸屬於『生活類』的新聞；但此篇新聞內容亦與香港政府相關，因此與『國際類』亦有若干關聯，若以一般傳統的單一類別的分類法，直接將此篇新聞歸屬於生活類，不但有欠周詳，也會造成使用者在『國際類』中找不到此篇報導的問題。因此，如果我們適度的給予此篇新聞模糊歸屬函數值，使其歸屬於『生活類』與『國際類』的程度分別為 0.9 與 0.7，則藉由模糊分類檢索功能，使用者不但可以在『生活類』中檢索出這篇新聞（如圖 4 所示），亦可在『國際類』中檢索到（如圖 5 所示），不但增加了檢索的彈性，也提升了回現率。



・新聞列表— Life (W / Fuzzy Relation)

日期	主題	
2002/3/18	公車處將加密平日上山班次	1.0
2002/3/18	今起台灣民眾赴港可辦電子簽證節省時間金錢	0.9
2002/2/25	大學學測 兩萬人未過關	0.9
2002/3/18	嘉義縣大尖山開發新景點	0.8

模糊類別：Life／相關係數>=0.7

圖 4 檢索『生活類』新聞之結果

・新聞列表— International (W / Fuzzy Relation)

日期	主題	
2002/4/3	馬里蘭校園 瘋狂到不行	1.0
2002/3/18	小鷹號航艦駛離日本橫須賀基地	1.0
2002/3/18	二億五千萬元 打造豪華派對	0.9
2002/2/25	印度廁所博物館 教你認識方便史	0.9
2002/3/18	六家台資銀行登陸 中共第二季審批	0.8
2002/3/18	新加坡出現大陸人留學和培訓新浪潮	0.8
2002/3/18	今起台灣民眾赴港可辦電子簽證節省時間金錢	0.7
2002/3/18	哈佛商學書刊在大陸暢銷	0.7
2002/3/18	大陸漁工兩週內解禁	0.7

模糊類別：International／相關係數>=0.7

圖 5 檢索『國際類』新聞之結果

此外，我們也在模糊分類檢索的功能中提供使用者自由選擇相關係數（即模糊歸屬函數值）之選項，以利使用者篩選新聞分類的相關度，相關係數愈高表示歸屬程度必須愈高者才會被檢索出來，如圖 6 所示。

圖 6 模糊分類檢索功能

(2) 語意檢索結果

本系統之新聞資料在經過 XML 標示加值處理之後，使用者除可依據標題、作者、時間 … 等資料欄位進行檢索外，也可針對新聞全文內容，依據人、事、時、地、物等語意條件進行更精確的內文語意檢索。例如，圖 7 乃是以「大學」作為關鍵字詞，選擇以不限標示的方式進行檢索之結果，共找到八篇文章；對於同樣的檢索詞，如果將內文檢索條件限制於「地」，表示使用者欲檢索之條件僅為內文語意上與「大學」相關的地方、地名或地點，而非所有與「大學」概念相關的文章，結果如圖 8 所示，找到七篇符合內文語意檢索條件之文章；再者，若檢索詞仍為「大學」，但將內文檢索條件限制於「事」，即表示使用者欲查詢之條件僅為有關大學的「事件」而非地點或其他，則如圖 9 所示，更精確的只找出兩篇在內文語意上含有大學相關事件的文章。

日期	主題	類別
2002/4/3	馬里蘭校園 瘋狂到不行	International
2002/4/3	大學生登山失蹤案 不排除謊報	Society
2002/4/3	泛藍軍有聲音 拱黃俊英選市長	Political
2002/3/18	哈佛商學書刊在大陸暢銷	Finance
2002/3/18	新加坡出現大陸人留學和培訓新浪潮	International
2002/3/18	保證留學大陸 補習班涉詐欺	Society
2002/2/25	大學學測 兩萬人未過關	Life
2002/2/25	三類官員 評價最差	Political

圖 7 不限檢索標示之結果

日期	主題	類別
2002/4/3	馬里蘭校園 瘋狂到不行	International
2002/4/3	大學生登山失蹤案 不排除謊報	Society
2002/4/3	泛藍軍有聲音 拱黃俊英選市長	Political
2002/3/18	哈佛商學書刊在大陸暢銷	Finance
2002/3/18	新加坡出現大陸人留學和培訓新浪潮	International
2002/3/18	保證留學大陸 補習班涉詐欺	Society
2002/2/25	三類官員 評價最差	Political

圖 8 限定檢索標示為「地」之結果

圖 9 限定檢索標示為「事」之結果

二、圖書模糊分類檢索系統

以下以一個簡單型式之實驗型研究範例說明之，在分類時給予二大部分的類號，第一明確適當的類號（A 部分），第二模糊類號（B 部分），以及模糊類號產生圖書之間的模糊關係（C 部分），並於模糊類號上給予每本書適當的模糊歸屬函數，使得一本特定圖書的主題就像地圖一樣能互通相連，讀者便能經由這個主題類號找到特定的圖書，也能在另一個類號裡找到相同之書籍，如一本書名為【圖書館學的哲學】，經由人為主觀性判斷屬於正規明確的類號屬於『1 類：哲學類』，經由人為模糊分類可在歸納分為『0 類：總類』，所以當讀者察找【圖書館學的哲學】這本書時，除了可在『0 類：總類』找到之外，也可以在『1 類：哲學類』中找到同一本。

以下圖 10 在作更進一步說明，所以由 B 與 C 部分可知，【圖書館學的哲學】這本書可從 C 部分判斷其相關書籍為【哲學與新聞】、【新聞心理學概論】、【土地倫理】，而關係度可從歸屬函數判斷其輕重，數值越接近於 1 越相似，反之毅然，因此一本書籍，假使其內容涵蓋多領域，其所屬類別也將分屬多類別，同類型書籍也將越廣。

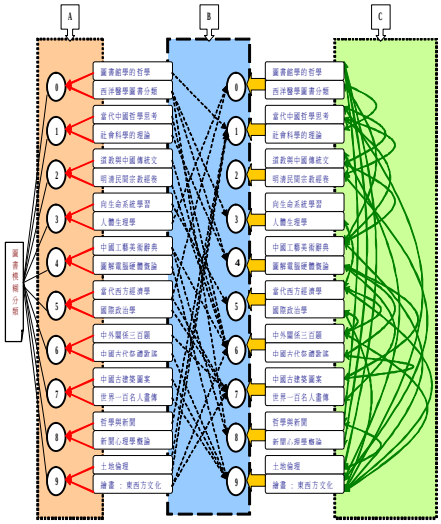


圖 10：精確分類與模糊分類關係圖

本研究設計了一個實驗型系統，其功能主要是在「模糊查詢部分」，此功能是透過模糊理論基礎概念加以應用於圖書分類法上，並加以整理給予每本圖書其歸屬函數值，來判別模糊程度高低，越接近於 1 代表越屬於某一類，反之毅然，因此透過此機制，可以將某書籍「同類」與「模糊類」相關書籍查找出來，藉以提高檢索回現率（Recall Rate）和精確率（Precision Rate）。

本研究設計了一個圖書模糊分類檢索系統，主要分為（A）前端使用模組：提供讀者透過個人需求選擇適當檢索方式進行書目檢索，依功能分為：(1)精確分類檢索模組（Precise_Search Module）、(2)模糊分類檢索模組（Fuzzy_Search Module ）、(3)相似書籍檢索模組（Similar_Search Module）（B）後端管理模組：提供系統管理者維護書目資料之新增、刪除、修改以及模糊分類度數選擇之動作，進入該模組前必須鍵入特定帳號與密碼提供系統驗證，才能進入進行管理動作。

依系統功能進一步說明：

（1）精確分類檢索模組（Precise_Search Module）提供使用者鍵入關鍵字並選擇以書名、作者、出版社、出版地、出版年、館藏地、ISBN 進行比對查詢。

（2）模糊分類檢索模組（Fuzzy_Search Module ）提供使用者選擇圖書模糊分類度數值，搭配圖書類別進行查詢，查詢結果之度數越高之書籍，表示越接近所選擇之圖書類別，代表越接近某類別。

（3）相似書籍檢索模組（Similar_Search Module）提供使用者選擇圖書類別，系統會將資料庫裡的書目資料，依照此類別的模糊度數高低全部呈列出來。

（4）後端管理模組（Black_Control Module）此模組進入前必須登入特定帳號與密碼，提供管理者維護書目資料新增、刪除、修改以及模糊分類度數選擇之異動等動作。

本實驗結果針對模糊分類檢索模組介面以及結果部份進行詳細說明，其檢索介面(如下圖 11、12)，類號部分提供使用者以下拉式方式選擇中國圖書分類法 0~9 類，模糊度部份也採用下拉表單方式供選擇其模糊歸屬函數值，以決定檢索資料時的精確度。並設計了一個中國圖書分類法明細項目（如圖 13），提供對圖書分類法架構不清楚的使用者，以便掌握圖書分類號，使檢索時所下的檢索策略更精確。

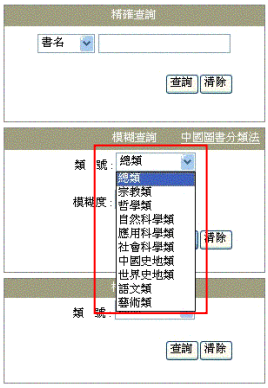


圖 11 模糊分類檢索模組-類號

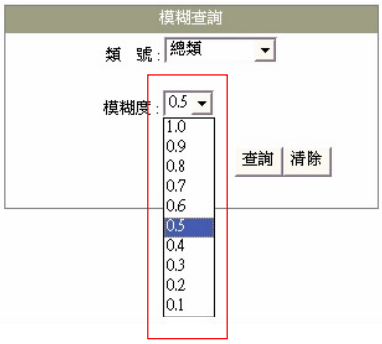


圖 12 模糊分類檢索模組-模糊度數

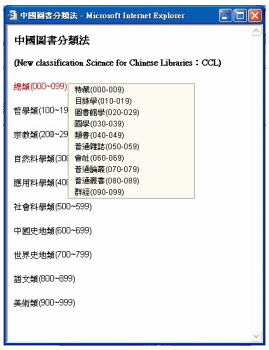


圖 13 中國圖書分類法

使用者選定類號與模糊度後會對系統內部資料庫書目資料進行比對，如核對成功系統會將書目資料依照所選擇類別之圖書模糊度數，依度數高至低呈列出來。如下圖 14：使用者類號選擇「總類」、模糊度選擇「0.1」，因此所有模糊度數高於 0.1 之總類圖書會將接近於 1 的某筆資料放於第一筆，以便使用者快速瀏覽。

從圖 14 中可知，「圖書館學的哲學」、「西洋醫學圖書分類系統析論」這二本書依內容來分類都含有「0 類：總類」之類別，表示使用者可以透過此模糊檢索之功檢索出同類型之書籍，由模糊度數來判斷「圖書館學的哲學」模糊度數 0.3 比「西洋醫學圖書分類系統析論」0.2 還接近於 1，因此較屬於總類，但兩者之間因度數相近，也代表兩者屬於同一類之書籍，所以藉此功能，能因而提昇圖書檢索上回現率（Recall Rate）和精確率（Precision Rate）更能減低使用者浪費在檢索上的時間。

書名	作者	出版社	出版地	出版年	ISBN	索書號	館藏地		C0
圖書館學的哲學	賴鼎銘	文華	台北市	1983	9578708033	020.1	中文圖書區		0.3
西洋醫學圖書分類系統析論	劉淑華	臺灣學生書局	臺北市	1979	957531582X	023.34	中文圖書區		0.2

圖 14 模糊分類檢索「0 類：總類」結果

如下圖 15 所示，檢索策略之類別為哲學類，模糊度數為 0.5，因此其檢索出的「人體生理學」、「哲學與新聞」、「新聞心理學概論」、「當代中國的哲學思考」、「圖書館學的哲學」五本書都屬於「1 類：哲學類」之書籍。因此本研究系統所設計的功能就在於模糊檢索介面裡提供了使用者自行選擇模糊度數（歸屬函數值）之選項，以利使用者依個人需求篩選出個人化資料。

登入號	書名	作者	出版社	出版地	出版年	ISBN	索書號	館藏地	C0	C1
10008	人體生理學	黃基礎	藝軒圖書出版社	臺北市	2002	9576165156	397 /8347	中文圖書區	0	0.9
10017	哲學與新聞	徐光春	新華書店	北京	1991	7200015814	890.1 /8564	中文圖書區	0	0.9
10018	新聞心理學概論	劉宗林	北京廣播學院出版社	北京	1993	7810040073	890.14 /8727	中文圖書區	0	0.9
10003	當代中國的哲學思考	肖明	經濟科學出版社	北京	1996	7505809598	120/855	中文圖書區	0	0.7
10001	圖書館學的哲學	賴鼎銘	文華	台北市	1983	9578708033	020.1	中文圖書區	0.3	0.6

圖 15 模糊分類檢索「1 類：哲學類」結果

肆、計畫成果自評

本研究在二年的過程中提出了一個結合模糊理論與全文標示之實際作法，並發展二套具有模糊分類搜尋與精確語意檢索的電子新聞管理系統與圖書模糊分類系統。其系統主要功能如下：

(一)電子新聞管理系統
(1)呈現出版模組
(2)資料檢索模組
(3)資料編輯模組
(4)資料管理模組
(二)圖書模糊分類檢索系統
(1)精確分類檢索模組
(2)模糊分類檢索模組
(3)相似書籍檢索模組
(4)後端管理模組

主要研究成果之一是藉由 Fuzzy 理論的技術，以歸屬函數的高低來判別各則新聞與圖書所屬類別輕重，並發展出具有模糊搜尋的檢索功能；另一個成果是藉由 XML 進行新聞內容的語意描述，各個進行資料處理及交換的模組皆以 XML 為基礎，系統內之所有資料亦採用 XML 格式，並發展出具有精確語意的檢索功能。由實驗結果可以清楚的看出結合此兩種不同作法，的確可以有效的提升回現率與精確率。

參考文獻

[1] 林信成、彭啟峰，*Oh！Fuzzy 模糊理論剖析*，台北：第三波，民 83。
[2] L. A. Zadeh, “Fuzzy sets,” *Information and*

Control **8** (1965), pp. 338-353.
[3] L. A. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. on Syst., Man and Cybern*. SMC-**3** (1973), pp. 28-44.
[4] Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall (1996).
[5] W3C, "Extensible Markup Language (XML)", available at <http://www.w3.org/XML/> (20 Feb. 2003).
[6] W3C, "Happy Fifth Birthday to XML-10 February 2003", available at <http://www.w3.org/> (20 Feb. 2003).
[7] W3C, "XML Schema", <http://www.w3.org/XML/Schema> (20 Feb. 2003).
[8] W3C, "Cascading Style Sheets", <http://www.w3.org/Style/CSS/> (20 Feb. 2003).
[9] W3C, "The Extensible Stylesheet Language (XSL)", <http://www.w3.org/Style/XSL/> (20 Feb. 2003).
[10] W3C, "Document Object Model (DOM) ", <http://www.w3.org/DOM/> (20 Feb. 2003).
[11] W3C, "Resource Description Framework (RDF) ", <http://www.w3.org/RDF/> (20 Feb. 2003).
[12] W3C, "Namespaces in XML", <http://www.w3.org/TR/REC-xml-names/> (20 Feb. 2003).
[13] W3C, "W3C Math Home", <http://www.w3.org/Math/> (20 Feb. 2003).
[14] W3C, "Scalable Vector Graphics (SVG)", <http://www.w3.org/Graphics/SVG/> (20 Feb. 2003).
[15] W3C, "PNG (Portable Network Graphics) ", <http://www.w3.org/Graphics/PNG/> (20 Feb. 2003).
[16] W3C, "Synchronized Multimedia", <http://www.w3.org/AudioVideo/> (20 Feb. 2003).
[17] XML.ORG, "Applying XML and Web Services Standards in Industry", <http://www.xml.org/ > (20 Feb. 2003).
[18] W3C, "Semantic Web", <http://www.w3.org/2001/sw/> (20 Feb. 2003).
[19] W3C, "Web-Ontology (WebOnt) Working Group", <http://www.w3.org/2001/sw/WebOnt/> (20 Feb. 2003).